Endometriosis risk alleles at 1p36.12 act through inverse regulation of *CDC42* and *LINC00339*

Joseph E Powell [1,2,*,+], Jenny N Fung [3,+], Konstantin Shakhbazov[2], Yadav Sapkota[3], Nicole Cloonan[3], Gibran Hemani[2,4], Kristine M Hillman[3], Susanne Kaufmann[3], Hien T Luong [3], Lisa Bowdler[3], Jodie N Painter[3], Sarah J Holdsworth-Carson[5], Peter M Visscher[2], Marcel E Dinger[6,7], Martin Healey[5], Dale R Nyholt[3,8], Juliet D French [3], Stacey L Edwards[3], Peter A W Rogers[5,+] and Grant W Montgomery[3,+]

1. The Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia

2. Centre for Neurogenetics and Statistical Genomics, Queensland Brain Institute, University of Queensland, St Lucia, Brisbane, Australia, 4072

3. Genetics and Computational Biology Department, QIMR Berghofer Medical Research Institute, Brisbane, Qld, Australia, 4006

4. MRC Integrative Epidemiology Unit, University of Bristol, Oakfield House, Bristol, BS8 2BN

5. Gynaecology Research Centre, University of Melbourne, Department of Obstetrics and Gynaecology, Royal Women's Hospital, Parkville VIC 3052, Australia

6. Garvan Medical Research Institute, Sydney, NSW 2010, Australia

7. St Vincent's Clinical School, University of New South Wales, Sydney, NSW 2052, Australia.

8. Institute of Health and Biomedical Innovation, Queensland University of Technology, Kelvin Grove, QLD 4059, Australia

+ Equal contribution

*** Corresponding author: joseph.powell@uq.edu.au**

ABSTRACT

Genome-wide association studies (GWAS) have identified markers within the *WNT4* region on chromosome 1p36.12 showing consistent and strong association with increasing endometriosis risk. Fine mapping using sequence and imputed genotype data has revealed strong candidates for the causal SNPs within these critical regions; however, the molecular pathogenesis of these SNPs is currently unknown. We used gene expression data collected from whole blood from 862 individuals and endometrial tissue from 136 individuals from independent populations of European descent to examine the mechanism underlying endometriosis susceptibility. Association mapping results from 7,090 individuals (2,594 cases and 4,496 controls) supported rs3820282 as the SNP with strongest association for endometriosis risk (p=1.84×10^{-5}, OR=1.244 (1.126-1.375)). SNP rs3820282 is a significant eQTL in whole blood decreasing expression of *LINC00339* (also known as *HSPC157*) and increasing expression of *CDC42* (p=2.0 x10^{-54} and 4.5x10^{-4} respectively). The largest effects were for two *LINC00339* probes (p=2.0 x10^{-54}; 1.0 x10^{-34}). The eQTL for *LINC00339* was also observed in endometrial tissue (p=2.4 x10^{-8}) with the same direction of effect for both whole blood and endometrial tissue. There was no evidence for eQTL effects for *WNT4*. Chromatin conformation capture provides evidence for risk SNPs interacting with the promoters of both *LINC00339* and *CDC4* and luciferase reporter assays suggest the risk SNP rs12038474 is located in a transcriptional silencer for *CDC42* and the risk allele increases expression of *CDC42.* However, no effect of rs3820282 was observed in *LINC00339* expression in Ishikawa cells. Taken together our results suggest that SNPs increasing endometriosis risk in this region act through CDC42, but further functional studies are required to rule out inverse regulation of both *LINC00339* and *CDC42*.

INTRODUCTION

Endometriosis is a common gynaecological disease, defined as the presence of endometrial tissue outside of the uterus in lesions containing endometrial glands and stroma[1]. It is associated with severe pelvic pain and infertility affecting 6-10%[2] of women during their reproductive years and 20-50% of women with infertility[3,4]. There is limited knowledge of the aetiology and pathogenesis, and accurate clinical diagnosis is usually by laparoscopy, an invasive and costly procedure.

Susceptibility to endometriosis is known to be influenced by genetic factors, with heritability of ~0.5 from twin studies[5]. Meta-analysis of two genome-wide association (GWA) studies[6,7] identified a genome-wide significant association for single nucleotide polymorphism (SNP) rs7521902 on 1p36.12 close to *WNT4*[6,8], a critical and well-known regulator of uterine development. Further GWA studies have been published that replicate the original association within the 1p36.12 region[9,10].

While GWA studies have proven to be powerful tools for identification of loci influencing disease susceptibility, the results have revealed little about the nature of the genetic component to these phenotypes. For endometriosis, fine mapping using sequence[11] and imputed genotype data[8] has identified the strongest signals within the WNT4 gene in the 1p36.12 region[11], however, causal variants and molecular pathogenesis are currently unknown.

One common mechanism by which GWA loci influence disease susceptibility is though mediating RNA transcription[12]. For most genes there are extensive inter-individual differences in RNA levels[13], and much of that variation is due to genetic factors[14-16]. Loci responsible for this transcription variation are termed expression Quantitative Trait Loci (eQTL), and their sentinel SNPs are often defined as eSNPs. Variation in genomic DNA can affect transcription in multiple ways. Most intuitively perhaps, eSNPs represent allelic variation in regulatory elements within the *cis*-region of transcripts that alter their expression level[17]. However, given high levels of localised linkage disequilibrium (LD), often

spanning multiple genes, the overlap of eQTLs and GWAS loci can be coincidental and not driven by the same functional variants.

In addition to the association with endometriosis, variation at the 1p36 region is also associated with bone mineral density (BMD) and a risk of ovarian cancer[18]. The primary BMD signal is located close to ZBTB40, and there is a secondary signal that overlaps with our association signal near WNT4[19]. The primary signal near ZBTB40 is correlated with reduced expression of WNT4 in fibroblasts, osteoblasts, and adipose tissue[19,20]. In ovarian cancer, the most strongly associated variant at the 1p36 locus is located in the promoter of WNT4[21]. Data from ovarian tumour cell lines identified a cis-eQTL for CDC42 for SNPs in the region[20,22]. Variants associated with endometriosis at 1p36 could act through one of several genes, and this may be tissue specific.

Here we present results from a study investigating gene expression data collected from whole blood and endometrial tissue from independent populations to identify eQTLs shared between blood and endometrial tissue for loci within 1p36.12. We show that the mechanism underlying endometriosis susceptibility does not act through regulation of the strong functional candidate *WNT4*, but through nearby genes, *LINC00339* (ENSG00000218510), currently annotated as a long non-coding RNA and Cell Division Control Protein 42 (*CDC42*) on chromosome one. We provide evidence of shared causal loci for SNPs increasing endometriosis risk and eQTLs controlling expression levels of *LINC00339* and *CDC42* in blood and *LINC00339* in endometrial tissue.

RESULTS

There is substantial evidence for genetic association with endometriosis susceptibility at chromosome 1p36.12 for a block of SNPs in high LD (Fig. 1) that spans ~130 kb and includes the genes *WNT4* and *CDC42*, and the non-coding RNA *LINC00339* (also known as *HSPC157*). In this study, we genotyped coding variants in all genes across the region in Australian cases and controls and combined the genotype data with previous GWAS results. We analysed 227 exome variants from the region around rs3820282 (+/- 2.25 Mb), and there was no evidence for association with any coding variants. In agreement with previous studies, the three SNPs showing the strongest association with endometriosis risk were rs3820282, rs56318008, and rs55938609 (**Table 1**). SNP rs3820282 [A/G] located at base-pair 22468215 in intron one of *WNT4* (p=1.84 x 10^{-5}, OR=1.24 (1.126-1.375)) is in strong LD (r^2>0.95, 1000 Genome CEU population) with the next two most significant SNPs (rs56318008 and rs55938609). Conditional analysis on rs3820282 including all polymorphic coding variants in *WNT4*, *CDC42*, *LINC00339*, and other genes in the region showed no evidence of additional independent signals.

Effect of endometriosis SNPs at 1p36.12 on RNA transcription in whole blood

Within the 1p36.12 locus, there are nine RefSeq genes that include one or more mRNA transcript probes assayed on the Illumina HT12-v4.0 array and expressed in RNA samples from whole blood in the Brisbane Systems Genetics Study (BSGS)[23]. These gene are; *C1QA*, *C1QB*, *C1QC*, *CDC42*, *EPHA8*, *LINC00339*, *LDLRAD2*, *USP48* and *ZBTB40*. After quality control (see Methods) there remain a total of 14 probes tagging transcripts of the nine genes within 1p36.12. Phenotypic correlations between the normalised expression levels show a low level of co-expression of transcripts within the 1p36.12 locus (**Supporting material figure 1**). Probes located within *WNT4* were not detected as expressed in the BSGS sample, in line with previous studies that fail to identify *WNT4* transcripts expressed in blood[24,25].

We initially investigated concurrence between the endometriosis fine-mapped sentinel SNP rs3820282 and eQTLs for these 14 probes (**Table 2**). The expression levels of three of the 14 probes in 1p36.12 show significant association with rs3820282 genotypes after correcting for multiple testing. The two probes with the largest effect (p=2.0 x10^{-54}; 1.0 x10^{-34}) were located in the long non-coding RNA *LINC00339* (lncRNA) (**Figure 1**). *LINC00339* is expressed in a wide range of healthy human tissues, including hematopoietic cells, ovaries, and uterus[24,25]. In blood, each copy of the risk allele (A) for endometriosis susceptibility of rs3820282 decreased the expression levels of the *LINC00339* probes ILMN_3272768 (ENST00000434233) by 0.86 standard deviations (SE 0.07) and ILMN_3194087 (ENST00000404210) by one standard deviation (SE 0.06) (Figure 2). SNP rs3820282 has a smaller, but still significant effect on the expression levels of ILMN_1675156 (ENST00000344548) in *CDC42* (P=4.4 x10^{-4}) (Table 2, Figures 1 and 2. For *CDC42*, each copy of the endometriosis risk allele (A), expression levels increased by 0.24 standard deviations (SE 0.07). The direction of the allelic effects of rs3820282 on whole blood expression of *LINC00339* and *CDC42* is consistent with previously reported results[26].

For the three probes located in 1p36.12 with significant eQTL effects for rs3820282, we performed a conditional analysis and identified secondary, independent eQTLs for all three probes (**Supporting material table 1**). While the eSNPs were different, no tertiary eQTL were detected after additional conditional analyses fitting each of the secondary eSNPs (see methods).

Expression of transcripts at 1p36.12 in endometrial tissue

Capture sequence of transcripts within this region demonstrate *WNT4*, *CDC42* and *LINC00339* are all expressed in endometrium with multiple transcripts (Montgomery and Shakhbazov unpublished). We analysed gene expression for probes in this region in endometrial samples from Illumina HT-12v4 expression arrays and after quality control, there were data for three probes for *LINC00339*, two probes for *CDC42* and one probe for *WNT4*. There was no evidence for effects of stage of the menstrual cycle on *LINC00339* expression. The two probes

for *CDC42* are located at 5' (ILMN_1675156) and 3' end (ILMN_1738424) of the gene. The *CDC42* probe at 3' end (ILMN_1738424) and the *WNT4* probe (ILMN_1666392) showed nominally significant evidence (p=0.013 and p=0.009 respectively) for differences across the menstrual cycle, where both *CDC42* and *WNT4* expression was highest in the early proliferative phase. The differences were not significant after correction for multiple testing. We found no evidence for differences in expression levels between endometriosis cases and controls for any assays after adjusting results for the stage of the cycle.

The effect of endometriosis-associated SNPs on RNA transcription in endometrial tissue is in the same direction as in blood.

For SNPs in the region typed on the Sequenom MassARRAY, effects of genotype on gene expression were tested after fitting stage of the cycle as a covariate. Expression levels for *LINC00339* probes (ILMN_1901198, ILMN_3194087, ILMN_3272768) all showed significant eQTLs with rs3820282 (p<7.4 x10^{-8}) (**Table 3, Figure 2**), with a comparable estimated effect size for each copy of the risk allele [A] of -0.55 in endometrial tissue. There were no significant effects of SNP genotypes on the expression of *CDC42* or *WNT4* in endometrial tissue. However, small differences in expression observed for *CDC42* probe ILMN_1675156 showed the same direction of effect as in RNA samples from whole blood with increased expression associated with the risk alleles for rs3820282 (**Figure 2**).

We also observed the secondary eQTL in endometrial tissue for *LINC00339* previously identified in whole blood RNA with rs12061255 (p=1.45 x10^{-9}; LD between rs10917120 and rs12061255 is r^2=1). The result remained significant after correcting for multiple testing. There was no evidence for the association between rs12061255 and endometriosis risk (p=0.7068) and no significant effects of this SNP on *CDC42* and *WNT4* expression in endometrium. Interestingly, the risk allele (minor allele, A) for the key endometriosis SNP rs3820282 was associated with a decrease in *LINC00339* expression, while the

minor allele (T) of SNP rs12061255 showed an increase in *LINC00339* expression.

While *LINC00339* is designated as a long non-coding RNA, its status is unclear as the sequence has a small open reading frame with a strongly predicted signal peptide and *N*-terminal trans-membrane domain. However, mapping of peptides, identified by mass spectrometry, in GM12878 and K562 cell lines[27] and kidney, urine and plasma[28] samples reveal no known translated proteins located within *LINC00339* coordinates. Sequence data identifies a four bp deletion in the second exon of the *LINC00339* transcript (rs3036899) and capture sequence data for *LINC00339* transcripts expressed in human endometrium (Montgomery and Shakhbazov unpublished) show all three alleles for this insertion/deletion variant located within the mRNA sequence. If translated, this deletion variant would result in a truncated protein without the trans-membrane domain. Our results show this four base-pair deletion in the second exon of *LINC00339* gene is in low LD with the sentinel SNP associated with endometriosis risk (rs3820282, r^2 =0.05). However, there is LD (r^2 =0.36) with the alternative SNP rs12061255 that shows a strong eQTL for *LINC00339,* but no association with endometriosis risk.

Evidence that the causal variants for expression and endometriosis risk are the same

One of the challenges arising from both GWAS and eQTL analyses is the precise identification of the disease-causing variant[29]. There was strong overlap between the local pattern of SNP effects on meta-analysis *p*-values for endometriosis risk and the blood expression *p*-values for the three transcripts (**Figure 1 and supporting material figure 2**). However, given the high levels of LD within this region, we used the Regulatory Trait Concordance (RTC) method of Nica *et al*.[30] to distinguish between shared loci and coincidental overlaps within the region. The RTC score ranges from 0 to 1, with values closer to 1 indicating shared causal regulatory effects. All three probes had high RTC scores with the rs3820282 eSNP (>=0.9) indicating strong evidence of shared loci between the

endometriosis GWAS loci and eQTLs for *LINC00339* and *CDC42* (**supporting material table 2**). The RTC scores of the secondary eQTLs were < 0.1 providing further evidence that only the rs3820282 eQTL is likely to have a role in endometriosis susceptibility. Because of the limited number of genotyped SNPs in the endometrial sample we were unable to perform the RTC analysis to test for the congruence of endometriosis causal loci and endometrial eQTL.

In addition to RTC we used a Summary-data-based Mendelian Randomization (SMR) method[31] to test further for the functionally relevant element(s) underlying the endometriosis GWAS loci at 1p36.12. SMR adopts a Mendelian Randomisation approach to test the functional association between the expression level of a gene (measured by probes) and a trait. We employed SMR to examine the association between eQTL data[23] for each of the 14 probes within the 1p36.12 region and endometriosis GWAS data[8]. At a Bonferroni adjusted *p*-value threshold (0.05/14) SMR identified significant associations between eQTLs for *LINC00339* and *CDC42* and the endometriosis GWAS loci at 1p36.12 (**Supporting material table 2**). Furthermore, we found no significant (p=1.5 x10^{-1}) association for a conditional analysis between the secondary eQTL for *LINC00339* (rs10917120) and endometriosis. However, it is important to note, that both RTC and SMR do not distinguish between causal relationships and pleiotropy.

The top risk SNPs fall within putative regulatory elements (PREs) that frequently interact with the *LINC00339* and *CDC42* promoter regions

Chromosome conformation capture (3C) was used to investigate chromatin interactions between the candidate target genes and risk-associated SNPs. The *LINC00339* promoter showed a strong interaction with a putative regulatory element (PRE1) located ~115 kb centromeric to the gene in Ishikawa cell lines (**Figure 3 and Supporting material figure 3**). The region spans the *WNT4* promoter and includes the top risk SNP rs3820282. An interaction was also detected between the *CDC42* promoter and another PRE (called PRE2) located in the first intron of *CDC42* (~24 kb centromeric to the promoter). PRE2 contains

SNP rs12038474, which showed a strong signal for endometriosis risk (p=1.73 x 10^{-4}, OR=1.21 (1.096-1.341)) and is in LD (r^2>0.77, 1000 Genome CEU population) with the top risk SNP rs3820282. This SNP also showed a significant eQTL for *CDC42* in BSGS whole blood samples (p=6.28 x10^{-10}). There was no evidence of interaction for any region or risk-associated SNP with *WNT4*. However, four SNPs close to the *WNT4* promoter could not be resolved by 3C (**Figure 3 and Supporting material figure 3**).

The regulatory capability of PRE1 and PRE2, combined with the effects of candidate SNPs, was further examined in luciferase reporter assays. For PRE1, inclusion of the reference or risk allele of SNP rs3820282 had no significant effect on the *LINC00339* promoter activity in Ishikawa cells (**Figure 3C**). On the evidence from Ishikawa cells rs3820282 is unlikely to act through transactivation of *LINC00339*. However, it is possible that rs3820282 affects chromatin looping between the PRE and *LINC00339*: which would not be observed in a luciferase reporter assay. In contrast, PRE2 constructs containing the reference allele of SNP rs12038474 reduced *CDC42* promoter activity, suggesting that PRE2 can act as a transcriptional silencer (**Figure 3C**). Consistent with the eQTL analysis, inclusion of the minor (risk-increasing) allele of the SNP significantly increased the *CDC42* promoter activity in Ishikawa cells. Given that SNP rs3820282 may alter an ESR1 binding site, we also examined the effects of estrogen induction on PRE1, but observed no additional effects (**Supporting material figure 4**).

DISCUSSION

Studies in endometriosis report strong evidence for genetic association with disease risk at chromosome 1p36.12[6,7] in an LD block that spans genes *WNT4*, *CDC42,* and *LINC00339* (also known as *HSPC157*). Conditional analyses for the strongest signal did not detect any evidence for additional signals in the region. Analysis of gene expression in whole blood[18,24] identified eQTLs for transcripts in this region with the strongest evidence for *LINC00339* and also evidence for eQTLs for *CDC42*[16,23]. *WNT4* is not expressed in samples from whole blood. Endometriosis risk alleles decreased expression of *LINC00339* and increased expression of *CDC42*. Results from a large meta-analysis of eQTL data from blood show a strong eQTL for *CDC42* (p=9.8 x10^{-198}) located directly over our signal for endometriosis risk (expression data for *LINC00339* probes were excluded during quality control of the data in this study)[26]. The signals for disease association and eQTLs completely overlap and regulatory trait concordance and summary-data-base Mendelian Randomisation methods provide strong evidence for shared causal regulatory effects.

The tissue(s) or cell types likely contributing to functional effects of endometriosis risk variants include viable endometrial tissue or endometrial stem cells deposited in the peritoneal cavity via retrograde menstruation[32-34]. We, therefore, analysed expression of *LINC00339*, *CDC42* and *WNT4* transcripts in RNA samples from endometrial tissue. Gene expression array results show that all three genes were expressed in the endometrium with evidence for differential expression of *CDC42* and *WNT4* during the menstrual cycle. *WNT4* is known to be critical for the development of the female reproductive tract[35] and the level of *WNT4* mRNA expression was significantly lower in human endometrial carcinomas than in the normal endometrium[36]. After adjusting for the stage of the cycle, we did not observe any significant effects of endometriosis risk alleles on *WNT4* expression in the endometrium.

There was strong evidence for eQTLs for *LINC00339* in endometrium for SNP rs3820282 with the same direction of effect as the eQTLs in blood. Effects of

rs3820282 on *CDC42* expression in the endometrium were not significant, but the direction of effect was similar to that observed in whole blood, where endometriosis risk alleles increased expression of *CDC42. LINC00339* and *CDC42* represent a complex locus with some evidence for combined *LINC00339*/*CDC42* transcripts in AceView[37]. The genes encode 22 different mRNAs, 18 alternatively spliced variants, and four unspliced forms with putative evidence for 15 spliced mRNAs encoding proteins. Results suggest variants affecting endometriosis risk at 1p36 affect expression of LINC00339 and CDC42 in opposite directions, but additional studies will be required to confirm a role for one or both genes in the pathogenesis of endometriosis or ovarian cancer.

Little is known about *LINC00339* function, despite near ubiquitous expression. Transcription at this locus is reported in pigs and cattle[38,39]. *LINC00339* expression appears to be inversely correlated with blood cholesterol levels; down-regulated in cells from patients with familial hypercholesterolemia[40], and up-regulated in patients with low baseline LDL levels[41]. Given that the risk allele reduced expression of *LINC00339*, these findings are consistent with significantly increased LDL levels in women with endometriosis[42] but do little to illuminate potential mechanisms of action. *LINC00339* was one of 108 cDNA clones identified by subtractive hybridization and up-regulated in endometriosis lesions compared with normal endometrium[43]. However, no subsequent studies have considered *LINC00339* expression in endometriosis or identified possible casual roles in endometriosis risk. In our results, the endometriosis risk allele for rs3820282 was associated with decreased expression of *LINC00339* in RNA samples from the endometrium and whole blood. We did not observe any differential expression of *LINC00339* across the menstrual cycle, but the endometriosis risk allele (A) for the sentinel SNP rs3820282 may alter an estrogen receptor (ESR1) binding site in several cell types[11] and *LINC00339* is over-expressed when the estrogen receptor is knocked down in MCF7 breast cancer cell lines[44].

Endometriosis is considered a benign disorder, but cells in endometrial implants have increased capacity to proliferate, implant and grow in the peritoneal

cavity[45]. *CDC42*, a member of the Rho family of GTPases, is known to act as a molecular switch that can activate some downstream targets[46]. It has been implicated in a variety of signalling cascades initiating changes in cellular processes including cell polarity, cytoskeleton remodelling, proliferation, migration, adhesion, membrane trafficking and transcription[47,48]. Increasing evidence has indicated that *CDC42* is involved in cell migration and tumour progression in multiple cancer types including hepatocellular carcinoma cells and colorectal cancer[49-51]. *CDC42* has been implicated in progression of both ovarian and breast cancer[52,53],

Expression of *CDC42* is reported[54] to be higher in ovarian endometriotic cysts compared to patients with adenomyosis suggesting increased expression of *CDC42* may contribute to the development of ovarian endometriosis. The key SNPs associated with endometriosis at the 1p36 locus are also strongly associated with risk for ovarian cancer[21] and endometriosis is a known risk factor for ovarian cancer[55] with the strongest evidence for genetic overlap with clear cell ovarian cancer[56]. Gene expression studies in high-grade serous ovarian cancer (HGSOC) samples from The Cancer Genome Atlas (TCGA) project demonstrated that risk SNPs for HGSOC at chromosome 1p36 were eSNPs for CDC42[22]. Elevated expression of CDC42 was associated with increased risk of HGSOC, and overexpression of the gene was associated with shorter population-doubling times and reduced migration of cells in culture. The functional data from 3C and luciferase reporter assays provide strong evidence that the SNP rs12038474, associated with endometriosis risk, influences promoter activity of *CDC42*. Taken together, results suggest *CDC42* plays an important role in the development of endometriosis and endometriomas, and progression to ovarian cancer in some patients.

We identified a secondary eQTL for *LINC00339* in the region in both whole blood and endometrial RNA samples. The second eSNP (rs12061255) is not associated with endometriosis risk, is in low LD with endometriosis-associated SNPs, and is not associated with up-regulation of *CDC42* expression. This is further supported by the larger study of expression in blood of Westra *et al.[26],* where there was no

evidence for eQTL effects of rs12061255 on *CDC42* expression[26]. The second eQTL for *LINC00339* argues against a causal role for LINC00339 acting alone. However, co-ordinated and inverse regulation of both *LINC00339* and *CDC42* expression by the causal variant(s) associated with endometriosis in this region may be important for SNP effects on endometriosis risk.

The association signals and critical SNPs for risk of endometriosis and ovarian cancer at this region completely overlap[8,21] and SNPs with the strongest association are located in the promoter region of *WNT4*. We saw no evidence for eQTL effects of rs3820282 on *WNT4* expression in endometrium and transfection of wild-type and risk haplotype *WNT4* promoter constructs into ovarian surface epithelial cells also had no significant effects on *WNT4* expression in luciferase reporter assays[21]. Candidate causal SNPs[21] with a likelihood of less that 1:100 for being causal in ovarian cancer span a region upstream of *CDC42* and across *WNT4* (chr1: 22,366,102-22,492,887). Some SNPs across the region are located in promoters and putative enhancers in relevant tissues[57] and alter transcription factor binding sites including HMGA1/FOXJ3/SOX13 (rs10917130), SMAD3 (rs3754496), BRCA1 (rs2268179), and ESR1 (rs3820282)[11,57].

In summary, this study of gene expression in whole blood and endometrium identified *LINC00339* as the gene with the strongest eQTL with risk alleles for SNPs associated with endometriosis at chromosome 1p36.12. There was evidence for smaller, but opposite effects on *CDC42* expression. The signals for disease association and eQTLs completely overlap and regulatory trait concordance and summary-data-base Mendelian Randomisation methods provide strong evidence for shared loci, although they are unable to distinguish between causal relationships and pleiotropy. Results from chromatin conformation capture show strong interactions between putative regulatory elements containing endometriosis associated SNPs and the promoters of *LINC00339* and *CDC42*. The Luciferase reporter assays support a direct effect of rs12038474 on expression of *CDC42*, but we do not observe direct effects of rs3820282 on *LINC00339* expression were in Ishikawa cells. Taken together, the

461 results strongly implicate variation in expression of *CDC42* in endometriosis risk.
462 Additional expression and functional studies will be necessary evaluate whether
463 there is any role for inverse regulation of *LINC00339* and *CDC42* to determine the
464 role(s) of *CDC42* in regulating risk of endometriosis and ovarian cancer.

METHODS

Genotyping and association analyses

A total of 2,213 surgically confirmed endometriosis cases and 2,044 controls were genotyped on HumanCoreExome chips (Illumina Inc, San Diego)[6,58]. Cases and controls for the HumanCoreExome genotyping included all Australian samples typed on Illumina 670-Quad (cases) and 610-Quad (controls) BeadChips (Illumina Inc)[6] for our previous genome-wide association study if DNA samples were still available. Genotype data across the chromosome 1 region for the same individuals were merged for Illumina I670/I610 data, HumanCoreExome data, and Sequenom MassARRAY custom genotypes[59] from a subset of samples including 930 of the surgically confirmed cases with a family history of endometriosis and a control group of 958 unrelated women (recruited for a study of twins who self-reported that they had never been diagnosed with endometriosis)[11]. Standard quality control procedures were applied to individual datasets as outlined previously[11]. Briefly, SNPs with >5% missing rate, out of Hardy-Weinberg Equilibrium ($p < 10^{-6}$) in controls and MAF < 1% were excluded. Samples with non-European ancestry or with low call rates (<95%) were excluded from the downstream analyses.

The final combined Australian dataset consisted of 2,594 cases and 4,496 controls. Of the total 7,090 individuals in the combined dataset, 6,503 are unrelated while 587 are related to some degree. The merged data was imputed using the MACH program[60,61] to impute missing genotypes. The quality of the imputed genotypes was assessed by R^2 metric, which estimates the squared correlation between true and imputed genotypes. All SNPs passed standard imputation quality control threshold (R^2 >0.3). Association analysis for markers across the 1p36.12 region was performed using an association analysis of imputed genotype dosage scores with endometriosis implemented through PLINK software (http://pngu.mgh.harvard.edu/purcell/plink/)[62]. To account for relatedness in the dataset, the analysis was conducted using a robust variance estimation approach[63,64] available in PLINK.

Gene expression in whole blood

We used gene expression data from the Brisbane Systems Genetics Study (BSGS) to investigate the effect of SNPs located within 1p36.12 on *cis*-located probes. BSGS comprises 862 individuals of European descent from 274 independent families[23]. DNA samples from each individual were genotyped on the Illumina 610-Quad Beadchip by the Scientific Services Division at deCODE Genetics Iceland. Full details of genotyping procedures are given elsewhere[23,65]. Filtered genotypes were then imputed to 1000 Genomes reference panel (release 3.0) using hapi-ur[66] and impute2[67]. SNPs with a poor imputation quality score (R^2 < 0.3) and with an MAF < 0.05 were removed.

Whole blood for expression profiling was collected directly into PAXgene tubes (QIAGEN, Valencia, CA). Total RNA was extracted from PAXgene tubes using the WB gene RNA purification kit (QIAGEN, Valencia, CA). RNA from all samples was run on an Agilent Bioanalyzer to assess RNA integrities and to estimate RNA concentrations. RNA was amplified and converted to biotinylated cRNA using the Ambion Illumina TotalPrep RNA Amplification Kit (Ambion).

Expression profiles were generated by hybridising 750 ng of cRNA to Illumina HumanHT-12 v4.0 Beadchips according to Illumina whole-genome gene expression direct hybridization assay Guide (Illumina Inc, San Diego, USA). Briefly, 500 ng of total RNA were used to generate biotinylated cRNA, which was fragmented and hybridised to an Illumina whole genome expression chip, HumanHT-12 v4.0 for 18 h at 58^{o}C. Beadchips were then washed and stained and subsequently scanned to obtain fluorescence intensities. Samples were scanned using an Illumina Bead Array Reader. Samples were randomised across chips and chip positions, with a check for balance across families, sex and generation.

The following normalisation procedures were applied to the raw expression data for the eQTL analysis. Pre-processing of data generated by the Illumina Bead Array Reader was done using Illumina software, GenomeStudio (Illumina Inc., San Diego). Pre-processing included; correction for chip background effects, removal of outlier beads, computation of average bead signal and calculation of detection *p*-values using negative controls present on the array. Removal of chip background effects can lead to negative expression levels for transcripts with low levels of expression. To avoid problems with further normalisation procedures, negative values were denoted as missing data identifiers. Thus, in subsequent normalisation procedures and analyses samples with probes coded as missing were ignored. Finally, we checked that probes did not contain sequence variants with MAF >0.05 in both the 1000Genomes and BSGS cohorts.

Quality Control

The Illumina HT-12 v4.0 chip contains 22 probes that tag transcripts located within the 1p36.12 region. To avoid spurious associations, we removed seven probes, mapping to five RefSeq genes that were not expressed in greater than 10% of samples. These genes were also not expressed in GTEx whole blood data[68]. Of the 14 probes remaining, the mean of the proportion of samples with *p*-values <0.05 was 97%, implying that relatively little missing data remained within the expression dataset.

Whole blood eQTL

The gene expression normalisation and eQTL mapping have been described in detail elsewhere[16,23]. However, we will briefly describe the methods here. To minimise the influence of overall signal levels, which may reflect RNA quantity and quality rather than a biological difference between individuals, the following standardisation procedures were applied. Adjusted expression levels for each probe were transformed using a Quantile transformation[69,70] to achieve a stabilized distribution across average expression levels. Further normalisation

was performed to allow expression levels to be compared across chips and genes. This was achieved by fitting the following linear mixed model;

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon} \qquad (1)$$

Where $\boldsymbol{y}$ is a vector of log-transformed probe expression levels, $\boldsymbol{\beta}$ is an unknown vector of fixed effects of the batch and blood cell counts (extraction date, gender, Red blood cells, platelets, neutrophils, monocytes, eosinophils, basophils, CD4, CD8, CD19, CD56). $\boldsymbol{\gamma}$ is an unknown vector of random effects of batch (chip and chip position, age) with known design matrix $\boldsymbol{Z}$, and $\boldsymbol{\varepsilon}$ is a vector of residual errors. The residuals from this model were standardised to z-scores and used in all further analyses.

The relationship between SNP genotypes and normalised probe expression levels had been tested for using a linear mixed model (--assoc command) implemented in MERLIN[71]. SNP genotype effects were estimated assuming an additive genetic model. A conditional analysis was used to address the potential of missing secondary eQTL in linkage disequilibrium (LD) with other eQTL. For each probe with an identified eQTL, we corrected for the main effects of the sentinel eSNP (SNP with the largest R^2) by regressing its genotypes against the expression levels. Residuals from this analysis were then used for the second round of eQTL mapping, allowing us to detect independent eQTL. If additional eQTL were identified from this second round of analysis, the process was repeated, correcting for the main effects of the top eSNP from the first and second eQTL using multivariate regression.

For probes with a significant association with rs3820282, we performed a conditional analysis fitting rs3820282 genotypes as a fixed covariate and testing for secondary effects on all SNP within +/- 1MB of rs3820282. The study-wide significance of secondary eQTL was determined as 0.05/ number of SNPs tested. If secondary eQTL were identified, we continued by fitting the genotypes of the secondary sentinel eSNP alongside rs3820282.

We used two methods to distinguish between shared causal effects and coincidental overlaps. The first is the Regulatory Trait Concordance (RTC) test, which is a rank-based score test that accounts for differences in the local LD structure between estimated eSNP and GWAS effects and is described in detail in Nica *et al*. [30]. The second is a Summary-data-based Mendelian Randomization (SMR) method, which tests the functional association between the expression level of a gene (measured by probes) and a trait through the regression of estimated effect sizes from the eQTL and GWAS analyses. Further details are provided in Zhu *et al*.[31].

Gene expression in endometrial tissue

Sample collection

Endometrial tissue samples were collected by curettage from 136 women recruited through the Royal Women's Hospital in Melbourne. Women undergoing laparoscopic surgery provided informed written consent before the operation. Only premenopausal women who were free from hormone treatment (in the three months prior to surgery) were included in this study. Detailed patient questionnaires, past and present clinical histories, pathology findings and surgical notes were recorded for each participant. A total of 93 women were surgically diagnosed with endometriosis by visual inspection at laparoscopy, 38 women had no history of endometriosis and a negative result at laparoscopy, and five were unknown because the surgical examination was inconclusive. Endometrial cycle stage was determined following histological assessment at pathology (4 Menstrual, 4 Early Proliferative, 58 Mid-Proliferative, 12 Late Proliferative, 16 Early Secretory, 24 Mid-Secretory, and 18 Late Secretory. Endometrial tissue samples were taken from the women by curette and were stored in RNA*later* (QIAGEN) at -80°C until RNA extraction. Whole blood from the same individuals was also collected to investigate the effect of SNPs located within 1p36.12 region on expression levels of *cis*-located transcripts in endometrial tissues. The study was approved by the Human Research Ethics

Committees of the Royal Women's Hospital in Melbourne and the QIMR Berghofer Medical Research Institute.

DNA was extracted from the whole blood and DNA samples were genotyped for a total of seven variants located within the 1p36.12 region (five top GWA/imputed SNPs, rs3820282, rs56318008, rs55938609, rs12037376, rs7521902, top eSNPs with LINC00339, rs12061255 and a 4-bp insertion/deletion variant, rs3036899) using the Sequenom MassARRAY technology (Sequenom Inc., San Diego, CA, USA). All SNPs had call rates >95%.

Total RNA was extracted from homogenized endometrial tissues using RNA lysis solution (RLT buffer) and RNeasy Plus Mini Kit according to the manufacturer's instructions (QIAGEN, Valencia, CA). RNA quality was assessed with the Agilent Bioanlayzer 2100 (Agilent Technologies, Santa Clara, CA), and concentrations were determined using the NanoDropND-6000. 250ng of RNA was amplified and converted to biotinylated cRNA using the Ambion Illumina TotalPrep RNA Amplification Kit (Ambion). Expression profiles in endometrial tissue were generated by hybridising 750 ng of cRNA to Illumina HumanHT-12 v4.0 Beadchips (as described above).

Endometrial tissue gene expression normalisation

Adjusted expression levels for each probe were transformed using a Quantile transformation[69,70] to achieve a stabilized distribution across average expression levels. Further normalisation was performed to allow expression levels to be compared across chips and genes. This was achieved fitting a linear mixed model as described above where $\boldsymbol{\gamma}$ is an unknown vector of random effects of the batch (chip and chip position) in endometrial tissue gene expression normalisation.

Logistic regression was used to test for differential gene expression between cases and controls and between phases of menstrual cycle of the tissue samples (proliferative and secretory phases determined from histological evaluation), with and without adjusting for phases of the menstrual cycle and case/control status, respectively. An interaction term in the logistic model was also included

to assess for possible interaction between phases of the menstrual cycle and case/control status.

eQTL analysis

For each of the seven variants examined, a *cis*-eQTL analysis was conducted to investigate the putative association between the variant and expression levels of nearby transcripts. The eQTL analysis was performed on the total of 123 tissue samples with recoded SNP genotypes based on minor allele dosage and fitted linear regression models, with phases of the menstrual cycle included as a covariate in the model. Study-wide significance was determined using a Bonferroni adjustment (0.05/number of tests performed).

Cell lines

The Ishikawa endometrial cancer cell line (kindly provided by Pamela Pollock, QUT, Brisbane) was grown in DMEM medium with 10% FCS and antibiotics. The cell line was maintained under standard conditions, routinely tested for *Mycoplasma* and short tandem repeat (STR) profiled to confirm cell line identity.

Chromosome conformation capture (3C)

3C libraries were generated using *Nco*I as described previously[72]. 3C interactions were quantitated by real-time PCR (qPCR) using primers designed within restriction fragments (**Supporting material table 3**). All qPCR was performed on a RotorGene 6000 using MyTaq HS DNA polymerase (Bioline) with the addition of 5 mM of Syto9, annealing temperature of 66^{o}C and extension of 30s. 3C analyses were performed in three independent 3C libraries with each experiment quantified in duplicate. BAC clones covering the 1p36 region were used to create artificial libraries of ligation products in order to normalize for PCR efficiency. Data were normalized to the signal from the BAC clone library and, between cell lines, by reference to a region within *GAPDH*. All qPCR products were electrophoresed on 2% agarose gels, gel purified and sequenced to verify the 3C product.

Plasmid construction and reporter assays

Promoter-driven luciferase reporter constructs were generated by insertion of DNA fragments (synthesized by GenScript) containing the *LINC00339* or *CDC42* promoters into the *Kpn*I and *Nhe*I sites of pGL3-Basic. A 1423 bp fragment containing the Putative Regulatory Element (PRE1) or a 2475 bp fragment containing the PRE2 were then cloned into *Bam*HI and *Sal*I sites of the modified pGL3-promoter constructs. The minor (risk-increasing) alleles of individual SNPs were introduced into the PRE sequences by mutagenesis (GenScript). Ishikawa cells were transfected with equimolar amounts of luciferase reporter plasmids and 50 ng of pRL-SV40 transfection control plasmid with Lipofectamine 2000. The total amount of transfected DNA was kept constant at 600 ng for each construct by the addition of pUC19 as a carrier plasmid. Luciferase activity was measured 24 hr post-transfection by the Dual-Glo Luciferase Assay System. To correct for any differences in transfection efficiency or cell lysate preparation, *Firefly* luciferase activity was normalized to *Renilla* luciferase, and the activity of each construct was measured relative to the promoter alone construct, which had a defined activity of 1. Statistical significance was tested by log transforming the data and performing 2-way ANOVA, followed by Dunnett's multiple comparisons test in GraphPad Prism.

Estrogen induction

Ishikawa cells were first incubated with 10nM Fulvestrant (ICI 182780, Sigma) for 48 hours, then transfected using Lipofectamine2000 and treated with either 100nM 17β-Estradiol (Sigma) or DMSO (vehicle control) for 24 hours. Luciferase activity was measured 24 hr post-transfection as described above. Quantitative PCR (qPCR) for the established estrogen-regulated gene *TFF1* was performed on Ishikawa total RNA extracted using Trizol (Life Technologies). qPCRs were performed on a RotorGene 6000 (Corbett Research) with a *TFF1* TaqMan assay (Hs00907239_m1) and normalized against *β-glucuronidase* (4326320E).

ACKNOWLEDGMENTS

We thank the cohort participants who contributed to these studies, research nurses Ranita Charitra, Tracy Middleton and Irene Bell who recruited and consented all the endometrial biopsy patients at the Royal Women's Hospital, and the surgeons and anaesthetists who collected tissue and blood samples. We thank the women who participated in the QIMR Berghofer Medical Research Institute study. The GWAS data were generated as part of a study supported by the Wellcome Trust (WT084766/Z/08/Z). Research reported in this publication was supported by National Health and Medical Research Council (NHMRC) project grants GNT1026033, GNT1049472, GNT1046880, GNT1050208, GNT1083405, and GNT1010374 and QIMR Berghofer seed funding grant. GWM and PMV are supported by NHMRC Fellowships (GNT1078399, GNT1078037). JEP is supported by an Australian Research Council DECRA (DE1310691). NC is supported by an Australian Research Council Future Fellowship (FT120100453).

References

1. Giudice, L.C. Clinical practice. Endometriosis. *N Engl J Med* **362**, 2389-98 (2010).
2. Gao, X. *et al.* Economic burden of endometriosis. *Fertility and Sterility* **86**, 1561-72 (2006).
3. Prevalence and anatomical distribution of endometriosis in women with selected gynaecological conditions: results from a multicentric Italian study. Gruppo italiano per lo studio dell'endometriosi. *Human Reproduction* **9**, 1158-62 (1994).
4. Meuleman, C. *et al.* High prevalence of endometriosis in infertile women with normal ovulation and normospermic partners. *Fertility and Sterility* **92**, 68-74 (2009).
5. Treloar, S.A., O'Connor, D.T., O'Connor, V.M. & Martin, N.G. Genetic influences on endometriosis in an Australian twin sample. *Fertility and Sterility* **71**, 701-10 (1999).
6. Painter, J.N. *et al.* Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. *Nat Genet* **43**, 51-4 (2011).
7. Uno, S. *et al.* A genome-wide association study identifies genetic variants in the CDKN2BAS locus associated with endometriosis in Japanese. *Nat Genet* **42**, 707-10 (2010).
8. Nyholt, D.R. *et al.* Genome-wide association meta-analysis identifies new endometriosis risk loci. *Nat Genet* **44**, 1355-9 (2012).
9. Albertsen, H.M., Chettier, R., Farrington, P. & Ward, K. Genome-Wide Association Study Link Novel Loci to Endometriosis. *PLoS One* **8**(2013).
10. Pagliardini, L. *et al.* An Italian association study and meta-analysis with previous GWAS confirm WNT4, CDKN2BAS and FN1 as the first identified susceptibility loci for endometriosis. *J Med Genet* **50**, 43-6 (2013).
11. Luong, H.T. *et al.* Fine mapping of variants associated with endometriosis in the WNT4 region on chromosome 1p36. *Int J Mol Epidemiol Genet* **4**, 193-206 (2013).
12. Nicolae, D.L. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet* **6**(2010).
13. Storey, J.D. *et al.* Gene-expression variation within and among human populations. *Am J Hum Genet* **80**, 502-509 (2007).
14. Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat Genet* **39**, 1202-7 (2007).
15. Price, A.L. *et al.* Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet* **7**, e1001317 (2011).
16. Powell, J.E. *et al.* Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. *PLoS Genet* **9**, e1003502 (2013).
17. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53 (2007).
18. Goode, L.L. *et al.* A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat Genet* **42**, 874-+ (2010).

19. Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat Genet* **44**, 491-501 (2012).
20. Rivadeneira, F. *et al.* Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat Genet* **41**, 1199-206 (2009).
21. Kuchenbaecker, K.B. *et al.* Identification of six new susceptibility loci for invasive epithelial ovarian cancer. *Nat Genet* **47**, 164-71 (2015).
22. Lawrenson, K. *et al.* Cis-eQTL analysis and functional validation of candidate susceptibility genes for high-grade serous ovarian cancer. *Nat Commun* **6**, 8234 (2015).
23. Powell, J.E. *et al.* The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. *PLoS One* **7**, e35430 (2012).
24. Liu, X., Yu, X., Zack, D.J., Zhu, H. & Qian, J. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics* **9**, 271 (2008).
25. Petryszak, R. *et al.* Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res* **42**, D926-32 (2014).
26. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-U195 (2013).
27. Consortium, E.P. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-40 (2004).
28. Farrah, T. *et al.* State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. *J Proteome Res* **13**, 60-75 (2014).
29. Cooper, G.M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**, 628-40 (2011).
30. Nica, A.C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* **6**, e1000895 (2010).
31. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).
32. Gargett, C.E. *et al.* Potential role of endometrial stem/progenitor cells in the pathogenesis of early-onset endometriosis. *Mol Hum Reprod* **20**, 591-8 (2014).
33. Gargett, C.E. & Masuda, H. Adult stem cells in the endometrium. *Mol Hum Reprod* **16**, 818-34 (2010).
34. Zeng, B. *et al.* Increased expression of importin13 in endometriosis and endometrial carcinoma. *Med Sci Monit* **18**, CR361-7 (2012).
35. Kobayashi, A. & Behringer, R.R. Developmental genetics of the female reproductive tract in mammals. *Nat Rev Genet* **4**, 969-80 (2003).
36. Bui, T.D., Zhang, L., Rees, M.C., Bicknell, R. & Harris, A.L. Expression and hormone regulation of Wnt2, 3, 4, 5a, 7a, 7b and 10b in normal human endometrium and endometrial carcinoma. *Br J Cancer* **75**, 1131-6 (1997).

37. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* **7 Suppl 1**, S12 1-14 (2006).
38. Uenishi, H. *et al.* PEDE (Pig EST Data Explorer): construction of a database for ESTs derived from porcine full-length cDNA libraries. *Nucleic Acids Res* **32**, D484-8 (2004).
39. Zimin, A.V. *et al.* A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol* **10**, R42 (2009).
40. Chen, H., Wang, L. & Jiang, J. Transcriptome and miRNA network analysis of familial hypercholesterolemia. *Int J Mol Med* **33**, 670-6 (2014).
41. Won, H.H. *et al.* Differentially expressed genes in human peripheral blood as potential markers for statin response. *J Mol Med (Berl)* **90**, 201-11 (2012).
42. Verit, F.F., Erel, O. & Celik, N. Serum paraoxonase-1 activity in women with endometriosis and its relationship with the stage of the disease. *Human Reproduction* **23**, 100-104 (2008).
43. Hu, W.P., Tay, S.K. & Zhao, Y. Endometriosis-specific genes identified by real-time reverse transcription-polymerase chain reaction expression profiling of endometriosis versus autologous uterine endometrium. *J Clin Endocrinol Metab* **91**, 228-38 (2006).
44. Al Saleh, S., Al Mulla, F. & Luqmani, Y.A. Estrogen receptor silencing induces epithelial to mesenchymal transition in human breast cancer cells. *PLoS One* **6**, e20610 (2011).
45. Sharpe-Timms, K.L. Basic research in endometriosis. *Obstet Gynecol Clin North Am* **24**, 269-90 (1997).
46. Stengel, K. & Zheng, Y. Cdc42 in oncogenic transformation, invasion, and tumorigenesis. *Cellular Signalling* **23**, 1415-1423 (2011).
47. Cerione, R.A. Cdc42: new roads to travel. *Trends Cell Biol* **14**, 127-32 (2004).
48. Aznar, S. & Lacal, J.C. Rho signals to cell growth and apoptosis. *Cancer Lett* **165**, 1-10 (2001).
49. Liu, M. *et al.* miR-137 targets Cdc42 expression, induces cell cycle G1 arrest and inhibits invasion in colorectal cancer cells. *Int J Cancer* **128**, 1269-79 (2011).
50. Ke, T.W. *et al.* MicroRNA-224 suppresses colorectal cancer cell migration by targeting Cdc42. *Dis Markers* **2014**, 617150 (2014).
51. Chen, Y.W. *et al.* p16 Stimulates CDC42-dependent migration of hepatocellular carcinoma cells. *PLoS One* **8**, e69389 (2013).
52. Zuo, Y., Wu, Y. & Chakraborty, C. Cdc42 negatively regulates intrinsic migration of highly aggressive breast cancer cells. *J Cell Physiol* **227**, 1399-407 (2012).
53. Bourguignon, L.Y., Gilad, E., Rothman, K. & Peyrollier, K. Hyaluronan-CD44 interaction with IQGAP1 promotes Cdc42 and ERK signaling, leading to actin binding, Elk-1/estrogen receptor transcriptional activation, and ovarian cancer progression. *J Biol Chem* **280**, 11961-72 (2005).
54. Goteri, G. *et al.* Expression of motility-related molecule Cdc42 in endometrial tissue in women with adenomyosis and ovarian endometriomata. *Fertility and Sterility* **86**, 559-65 (2006).

55. Sainz de la Cuesta, R. *et al.* Histologic transformation of benign endometriosis to early epithelial ovarian cancer. *Gynecol Oncol* **60**, 238-44 (1996).
56. Lu, Y. *et al.* Shared genetics underlying epidemiological association between endometriosis and ovarian cancer. *Hum Mol Genet* (2015).
57. Coetzee, S.G. *et al.* Cell-type-specific enrichment of risk-associated regulatory elements at ovarian cancer susceptibility loci. *Hum Mol Genet* **24**, 3595-607 (2015).
58. Nyholt, D.R. *et al.* Genome-wide association meta-analysis identifies new endometriosis risk loci. *Nat Genet* (2012).
59. Zhao, Z.Z. *et al.* KRAS variation and risk of endometriosis. *Molecular Human Reproduction* **12**, 671-6 (2006).
60. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406 (2009).
61. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-34 (2010).
62. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559-75 (2007).
63. Barlow, W.E. Robust variance estimation for the case-cohort design. *Biometrics* **50**, 1064-72 (1994).
64. Williams, R.L. A note on robust variance estimation for cluster-correlated data. *Biometrics* **56**, 645-6 (2000).
65. Medland, S.E. *et al.* Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am J Hum Genet* **85**, 750-5 (2009).
66. Williams, A.L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *Am J Hum Genet* **91**, 238-51 (2012).
67. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
68. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
69. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93 (2003).
70. Smyth, G.K. & Speed, T. Normalization of cDNA microarray data. *Methods* **31**, 265-73 (2003).
71. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101 (2002).
72. Ghoussaini, M. *et al.* Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nat Commun* **4**, 4999 (2014).

FIGURES

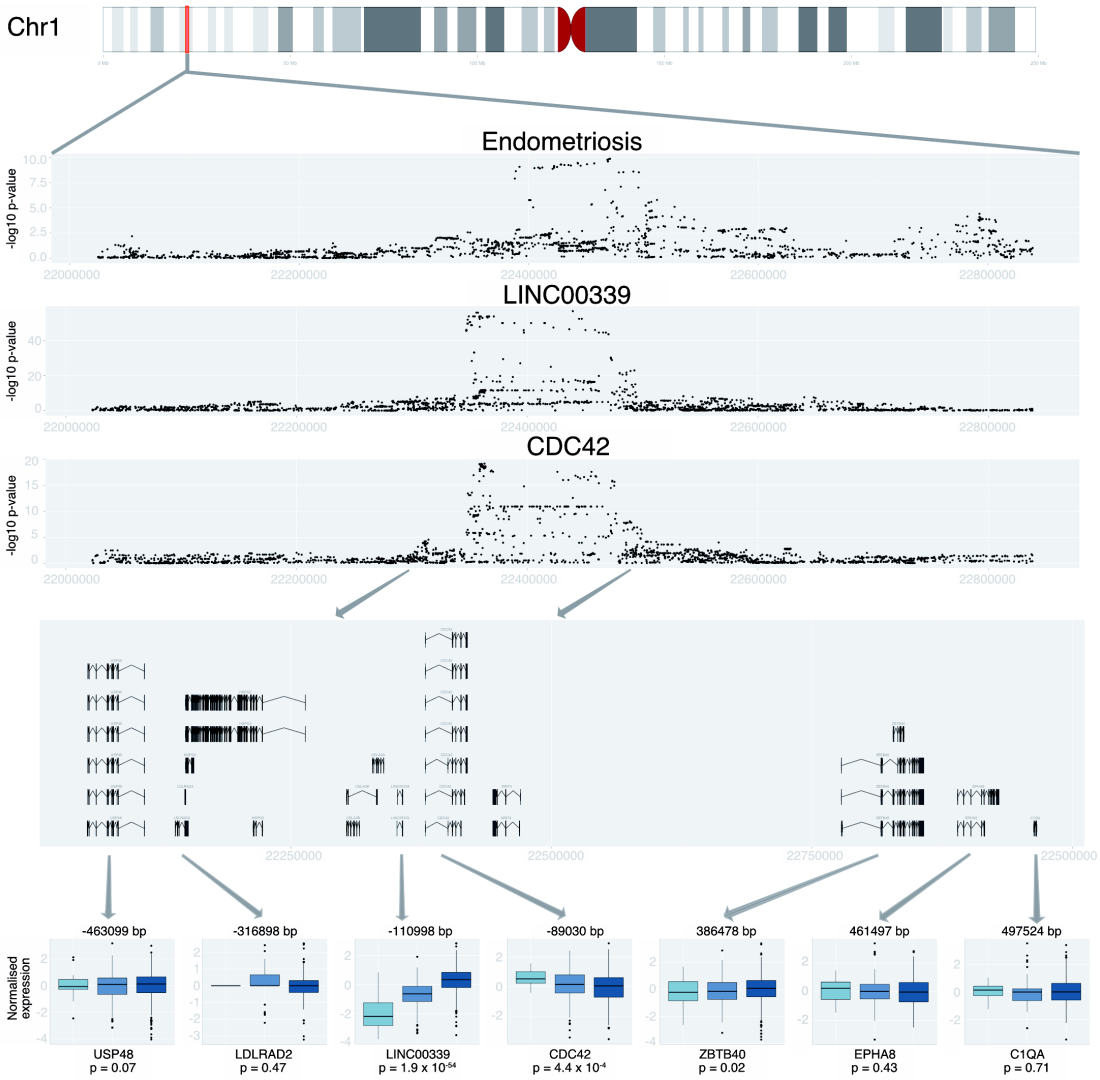

Figure 1 | SNPs within the 1p36.12 region are associated with both endometriosis risk, and the expression levels of LINC00339 and *CDC42*. Association results for individual SNPs are plotted by position on chromosome 1 (X-axis) as –log10 *p*-values (Y-axis) for endometriosis risk (first panel), *LINC00339* (second panel) and CDC42 (third panel) expression in the BSGS. The relative locations of genes within the 1p36.12 regions are shown in panel 4. The fifth panel shows the relationship between rs3820282 genotypes and transcript expression levels in BSGS for seven genes in the 1p36.12 region.

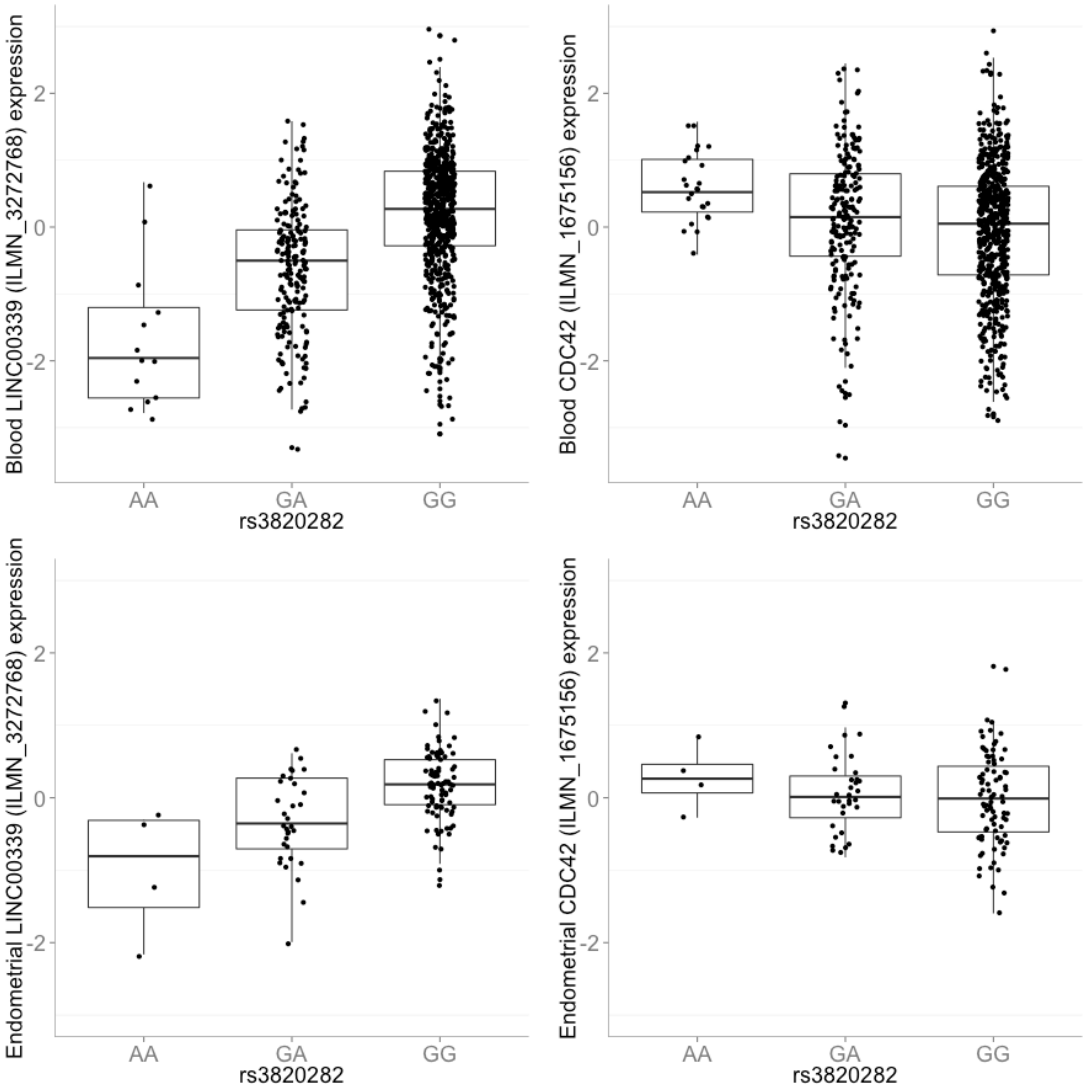

Figure 2 | Expression levels of *LINC00339* (left hand panels) and *CDC42* (right hand panels) with eQTL effects for the relationship between rs3820882 and gene expression in blood (top panel); rs3820282 and gene expression in endometrium (bottom panel). The overall estimated effect on *LINC00339* (ILMN_3272768) of each additional copy of rs3820282 endometriosis risk allele [A] is -0.86 in whole blood and -0.55 in endometrial tissue. The corresponding effect in CDC42 (ILMN_1675156) is 0.24 and 0.13 respectively.

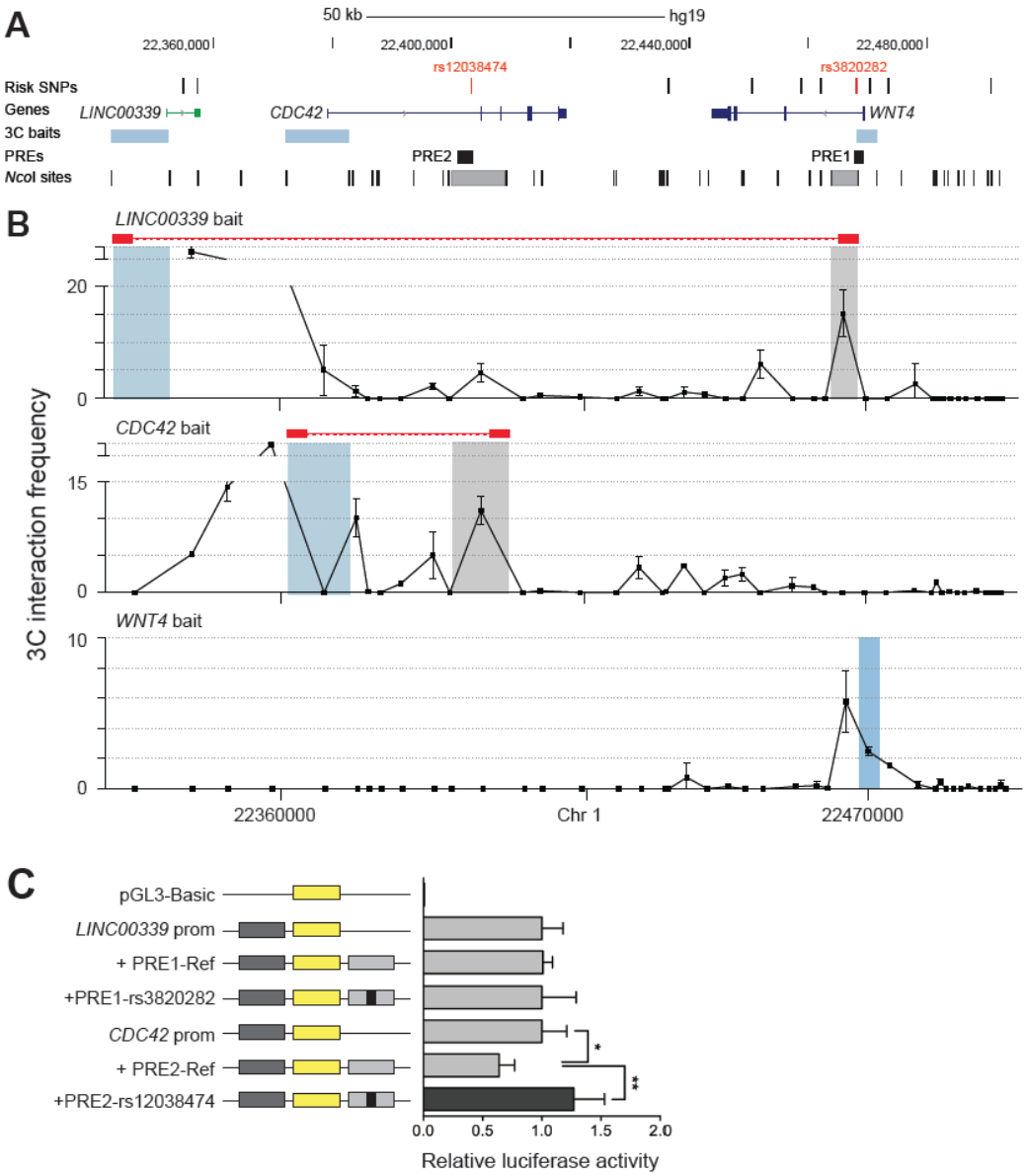

Figure 3 | Candidate causal SNPs are located within PREs that interact with the *LIN00339* or *CDC42* promoters. (a) The location of candidate causal SNPs are represented by black or red ticks, gene structures are depicted with exons (vertical boxes) joined by introns (lines). 3C anchors are shown as blue boxes, frequently interacting *Nco*I fragments as grey boxes and Putative Regulatory Elements (PRE1, PRE2) as black boxes. **(b)** 3C interaction profiles between *LINC00339*, *CDC42* or *WNT4* promoters and the 1p36 risk region in Ishikawa cell lines. 3C anchors are shown as blue boxes and frequent interactions highlighted with red connecting bars. Graphs represent one of three biological replicates. Error bars represent SD. **(c)** Luciferase reporter assays in Ishikawa cells. PREs containing the major SNP alleles (Ref) were cloned downstream of target gene promoter-driven constructs. Minor (risk-increasing) SNP alleles were engineered into the constructs and are designated by the rs ID of the corresponding SNP. Error bars denote 95% confidence intervals from three independent experiments. *P*-values were determined by 2-way ANOVA followed by Dunnett's multiple comparisons test (*P<0.05, **P<0.01).

Table 1 | Endometriosis association information for common SNPs with the key SNP (rs3820282) and significant association with endometriosis risk ($P<5\times10^{-3}$). Fine mapping results were from 7,090 individuals (2,594 cases and 4,496 controls) in the combined Australian dataset.

SNPs	Position(hg19)	LD with rs3820282 (r^2)	RA	OA	RAF#case	RAF#control	OR*	*P*
rs3820282	22468215	1	A	G	0.189	0.162	1.244(1.126-1.375)	1.84×10^{-5}
rs56318008	22470407	0.95	T	C	0.186	0.159	1.243(1.125-1.374)	2.06×10^{-5}
rs55938609	22470451	0.95	C	G	0.186	0.159	1.240(1.121-1.372)	2.88×10^{-5}
rs12037376	22462111	0.90	A	G	0.193	0.167	1.229(1.113-1.356)	4.24×10^{-5}
rs2235529	22450487	0.84	A	G	0.188	0.161	1.216(1.107-1.335)	4.61×10^{-5}
rs12404660	22458794	0.79	G	A	0.222	0.195	1.216(1.107-1.336)	4.71×10^{-5}
rs7412010	22436446	0.90	C	G	0.196	0.170	1.211(1.100-1.333)	9.71×10^{-5}
rs7515106	22473410	0.61	C	T	0.240	0.213	1.174(1.077-1.280)	2.66×10^{-4}
rs2473295	22354866	0.05	G	A	0.771	0.748	1.156(1.058-1.263)	1.39×10^{-3}
rs760923	22357217	0.05	T	G	0.770	0.748	1.154(1.057-1.261)	1.44×10^{-3}
rs7521902	22490724	0.61	A	C	0.264	0.240	1.138(1.048-1.236)	2.08×10^{-3}

Risk allele frequency
* Odd ratios were calculated for the risk allele

Table 2 | Effect of the fine-mapped (rs3820282) and original GWA sentinel (rs7521902) endometriosis SNPs on the expression levels of transcripts located within 1p36.12 locus. Expression levels of probes were measured in whole blood for 862 individuals from the Brisbane Systems Genetics Study[23].

			-Log10(p-value)		**Effect (SE)**	
Gene	**Probe**	**Probe start (bp)**	**rs3820282 [A]**	**rs7521902 [A]**	**rs3820282 [A]**	**rs7521902 [A]**
USP48	ILMN_2285141	22005116	$7.1x10^{-2}$	$6.1x10^{-1}$	-0.07 (0.07)	-0.03 (0.06)
USP48	ILMN_1756873	22005253	$3.8x10^{-1}$	$7.4x10^{-1}$	-0.06 (0.07)	-0.02(0.06)
USP48	ILMN_1738572	22054538	$7.4x10^{-1}$	$8.7x10^{-1}$	0.02 (0.07)	0.01 (0.06)
USP48	ILMN_1777726	22055186	$7.4x10^{-1}$	$8.1x10^{-2}$	-0.03 (0.10)	0.16 (0.09)
LDLRAD2	ILMN_1734125	22151317	$4.7x10^{-1}$	$1.5x10^{-1}$	-0.14 (0.19)	0.22 (0.15)
LINC00339	ILMN_3194087	22357217	**$1.9x10^{-54}$**	**$1.2x10^{-18}$**	**-1.00 (0.06)**	**-0.49 (0.05)**
LINC00339	ILMN_3272768	22357427	**$1.0x10^{-34}$**	**$2.5x10^{-14}$**	**-0.86 (0.07)**	**-0.45 (0.06)**
CDC42	ILMN_1675156	22379185	**$4.5x10^{-4}$**	$4.9x10^{-2}$	**0.24 (0.07)**	0.11 (0.06)
CDC42	ILMN_1738424	22419001	$1.3x10^{-1}$	$3.7x10^{-1}$	0.10 (0.07)	0.05 (0.06)
ZBTB40	ILMN_1784037	22854693	$2.4x10^{-2}$	$1.2x10^{-1}$	-0.15 (0.07)	-0.09 (0.06)
EPHA8	ILMN_1756989	22929712	$4.3x10^{-1}$	$6.6x10^{-1}$	-0.06 (0.07)	0.03 (0.06)
C1QA	ILMN_1737918	22965739	$7.1x10^{-1}$	$5.2x10^{-1}$	0.05 (0.12)	-0.07 (0.10)
C1QC	ILMN_1785902	22974217	$6.1x10^{-1}$	$6.3x10^{-1}$	-0.09 (0.17)	-0.07 (0.14)
C1QB	ILMN_1796409	22987790	$6.1x10^{-1}$	$9.7x10^{-1}$	0.08 (0.16)	-0.01 (0.13)

Table 3 | SNP effects for transcription levels measured in endometrial tissue.

						WNT4 (ILMN_1666392)			*CDC42* (ILMN_1675156)			*LINC00339* (ILMN_1901198)		
SNP	**Position (hg19)**	**LD*** **(r^2)**	**Endo -log10(P)**	**RA**	**OA**	**Effect of RA**	**-log10(P)**	**Dist from SNP(bp)**	**Effect of RA**	**-log10(P)**	**Dist from SNP(bp)**	**Effect of RA**	**-log10(P)**	**Dist from SNP(bp)**
rs3820282	22468215	1	$1.8x10^{-5}$	A	G	0.08	$5.6x10^{-1}$	21735	0.07	$5.4x10^{-1}$	89030	-0.52	$2.3x10^{-8}$	110828
rs56318008	22470407	0.95	$2.1x10^{-5}$	T	C	0.15	$3.1x10^{-1}$	23926	0.03	$7.7x10^{-1}$	91221	-0.47	$1.5x10^{-6}$	113019
rs55938609	22470451	0.95	$2.9x10^{-5}$	C	G	0.14	$3.3x10^{-1}$	23971	0.03	$7.5x10^{-1}$	91266	-0.47	$1.8x10^{-6}$	113064
rs12037376	22462111	0.90	$4.3x10^{-5}$	A	G	0.09	$5.2x10^{-1}$	15630	0.07	$5.3x10^{-1}$	82925	-0.53	$2.2x10^{-8}$	104723
rs7521902	22490724	0.61	$2.1x10^{-3}$	A	C	0.20	$9.7x10^{-2}$	44243	0.004	$9.5x10^{-1}$	111538	-0.35	$4.4x10^{-5}$	133336
rs12061255	22350297	0.07	$7.1x10^{-1}$	T	C	0.05	$6.4x10^{-1}$	-96315	0.05	$5.9x10^{-1}$	-29020	0.47	$1.4x10^{-9}$	-7222
rs3036899	22357435	0.05	-	TCTT	-	0.07	$4.8x10^{-1}$	-89045	0.009	$9.1x10^{-1}$	-21750	0.21	$4.2x10^{-3}$	48

* LD (r^2) with the sentinel rs3820282 for association with endometriosis risk